基于舆情数据中台的产品多元化体系建设

——以南方舆情为例

摘　要：随着舆情市场的不断开拓，如何灵活响应多变性、多样化的用户需求，快速生成多元化产品服务，成为重要问题。本文立足舆情应用场景，通过标准的规范定义和服务的封装编排，构建一个承接技术、引领业务，可快速连接萃取的智慧数据中台，高效满足前台的数据分析和产品服务，引领舆情业务向纵深层次发展。

关键词：舆情；数据中台；数据建模
中图分类号：TP393　　文献标识码：A
文章编号：1671-0134（2019）01-119-03　　DOI：10.19483/j.cnki.11-4653/n.2019.01.033

文 / 吴娴　肖卓明　洪丹

引言

近年来，传统媒体不断寻求融合转型之道，拓展“媒体+”服务，为用户创造更多价值。为构建舆论引导新格局，越来越多的传统媒体整合品牌资源、政经资源和信息资源，切入舆情服务领域。

随着舆情市场的不断开拓，政务用户和企业用户之间、省级政务用户与区县基层政务用户之间，甚至地方政府用户和职能厅局用户之间，对舆情产品服务呈现多样化需求，同一用户在不同环境下对舆情管理的需求也相当多变。当这种变量积累到一定体量，为每个用户的定制开发成本会非常高，同时出现产品效率不高等问题。本文从南方舆情的实际业务发展出发，学习实践阿里巴巴首提的“大中台、小前台”概念，引入舆情数据中台的运转思路，支撑产品应用多元化快速生成，打造一揽子舆情产品服务，通过“技术降本、应用提效、业务赋能”，抓住舆情市场的发展机遇。

1. 难题与挑战

面对复杂的舆情应用场景，突破传统的系统架构，构建舆情数据中台，贴近用户多变多样的使用需求，面临着诸多技术难题与挑战。

挑战一：全域数据采集与入库。以需求为驱动，如何实时采集和引入多渠道数据（网站、论坛、博客、APP、微博、微信公众号、电台电视台）、多形态（自身业务系统、互联网采集、第三方交换）的数据，构建多信源、海量和动态的基础数据池存在很大的挑战。

挑战二：规范数据架构与研发。如何构建数据的分层与水平解耦结构，通过全域采集数据格式的规范化、交互接口的标准化实现架构的统一性、可靠性和灵活性，快速支撑上层数据应用和服务，是一个值得探讨的技术难点。

挑战三：跨域数据整合与知识沉淀。如何建立融合模型，通过不同维度的建模实现跨域舆情数据的整合，同时挖掘舆情数据从个体标签化到全局指标化，深度萃取数据价值，实现共性应用的知识沉淀，是面向舆情业务支撑提供底座能力的关键。

挑战四：数据封装应用与服务开放。数据的规模化发展是提供服务化能力。如何按应用要求做服务的封装，通过多元化的产品形态开放给外部服务用户，实现数据价值的快速分享，打通服务用户的最后一公里，是建设舆情数据中台的最终目的。

2. 技术架构与关键技术

数据中台的概念首先由阿里巴巴提出，“构建规范定义的、全域可连接萃取的、智慧的数据处理平台”，其建设目标是高效满足前台数据分析和应用的需求。为应对舆情服务需求的复杂多变性，南方舆情从实际业务出发，设计和搭建了舆情数据中台，以期实现产品定制化、服务个性化的快速部署。总体架构和关键技术描述如图 1 所示。

2.1 舆情数据采集：全域数据智能采集与入库

全域数据智能化采集平台主要对接的数据形态包括互联网数据采集、合作互补数据、媒体独家线下信源、自身采编业务数据。互联网数据通过分布式爬虫、智能采集调度、自适应采集策略、数据采集代理、自动登录验证等技术，灵活配置采集规则、抓取深度、扫描频率等采集策略，实现各渠道数据源的统一采集管理。依托分布式架构、多点负载均衡和自适应带宽设计，确保实时采集效率、采集稳定性和采集数据完整性。

以分布式计算架构实现对大规模数据的快速识别与信息抓取，对不同的信息使用不同的抓取策略，实现互联网信息抓取的自动化。采用分布式多线程并发指令执行体系结构、增量实时索引、智能分词等技术，采集和数据管理效率高。实现多个网站同时并发抓取、一个任

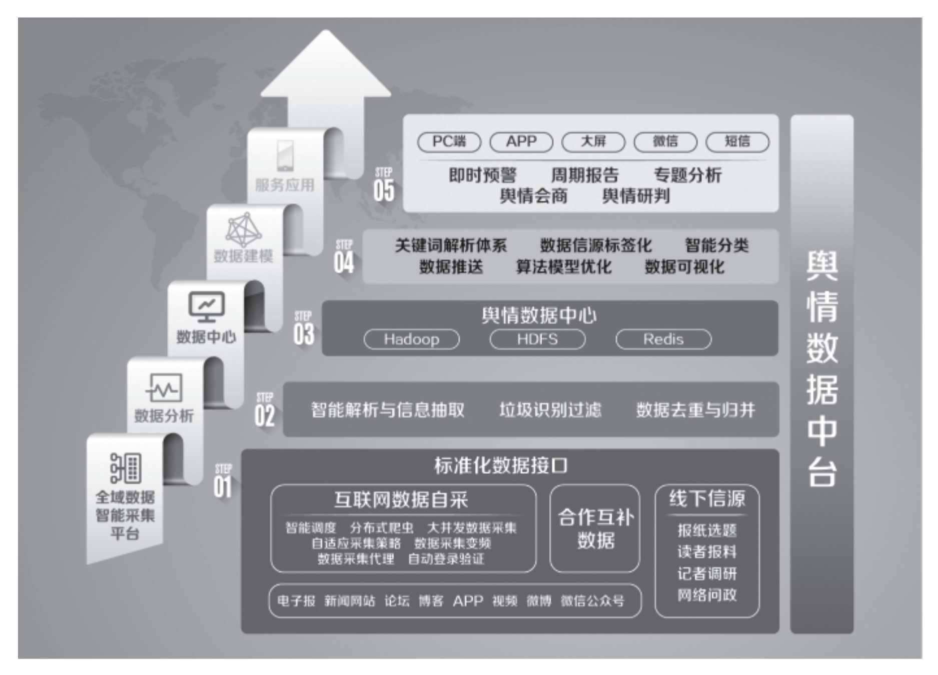


图 1　舆情数据中台的技术架构

务分布式并发多点处理、多点负载均衡的效果，可以防止短时间内向同一个网站发送过多的访问请求，提高大数据采集的效率和性能。运用 IP 代理池以及 API 模仿机制，对高频更新的数据进行 IP 轮询采集，能有效防止站点对系统 IP 的限制，同时系统能智能主动降低采集频率，降低 IP 被封的可能性。分布式采集的智能化调度，能有效提高数据采集的稳定性。

自采的互联网数据、合作互补数据、线下信源数据、采编业务数据经过标准化数据接口统一格式后进入数据分析层，打通数据孤岛，解决舆情数据的多源异构问题，减少烟囱式协作，确保舆情数据的多元性和完整性。

2. 2 舆情数据分析：数据标准规范化和可获取性

对采集到的信息进行垃圾识别过滤，自动清洗广告、无关图片、超链接、动态 Flash 等无用信息，利用智能解析，自动抽取标题、时间、来源、作者、正文等有效信息要素，通过内容判重引擎，根据数据内容分析语义对数据进行去重与归并，自动判断重复文章，实现自动去重与合并。采用分布式存储集群对加工后的标准化舆情数据、快照、索引进行存储，实现结构化、非结构化数据资源的融合管理。分析处理后的标准规范化舆情数据，为舆情服务应用提供调用基础，通过服务接口响应舆情业务的基本需求。

2. 3 舆情数据建模：数据多维标签与指标化

对海量舆情数据进行深入挖掘，利用关键词正则表达式智能解析匹配提取事件关联信息，并针对热点事件信息进行多维度分析。对事件信息进行分词、情感分析、热度分析、高频词提取、关联分析、数据统计等处理，结合自动摘要、分类、聚类等智能化运算，从而分析得出事件的发展趋势、敏感指数、地域分布、传播路径、关键人物、正负面倾向、网民观点等，深入分析事件的本质原因，形成建模基础数据。

以智能化标签的方式对数据信源进行归类，在逻辑上将数据信源自由组合成任意不同的虚拟数据信源包。在数据检索时，既可以在全局数据信源里进行匹配，也可以根据不同用户的不同需求，在虚拟数据信源包里进行数据匹配，缩小数据检索范围，提高数据检索精度，同时提高数据检索效率，实现舆情数据检索的灵活部署，快速响应业务环境的变化对业务流程优化提出的要求，为个性化、定制化的舆情产品提供基本支撑。

2. 4 舆情服务应用：数据应用封装与服务开放

利用与（+）、或（|）、非（–）无限层级优先级嵌套匹配规则，基于高效索引和排序算法的多维度检索实现关键字解析体系，支持多种索引条件的复杂组合，最大程度满足各种数据应用的封装需求。通过自动推送脚本将检索结果进行智能推送，便于舆情数据的高效共享，为进一步的舆情业务和其他舆情扩展业务发展提供强大的数据支撑，实现了一次跟踪，多端使用。推送使用 XML Schema 规范作为数据交换的标准格式，屏蔽了异构数据源之间的差异；数据格式采用 XML/JSON，方便调用，适配性强。

在舆情数据建模的基础上，通过虚拟数据信源包与关键词解析体系，对数据进行封装，结合智能推送开发多种舆情服务应用，譬如即时预警、周期报告、专题分析、舆情会商、舆情研判等，利用 PC 端、APP、大屏、微信、短信等多种发布渠道，形成舆情服务应用矩阵，满足全方位的舆情服务开放。

3. 应用案例

南方舆情通过应用创新和技术创新，基于舆情数据

中台构建产品快速生成的服务体系，以下简要阐述舆情数据中台实现业务赋能的落地应用案例。

3.1 社情风险指数和榜单

社情风险指数是南方舆情基于“数据沉淀、业务下沉”的特色产品应用。该产品生成逻辑和技术实现步骤如下：

第一，采集汇总历史风险事件，形成以业务核心对象为中心的连接和标签体系，并对风险事件性质、等级及传播范围提取要素，并对各要素进行赋值定义，建立社情风险指数计算模型；第二，一定周期内（日、周、月、年），增量舆情数据与离线历史数据同步共享，基于数据标准和标签模型开展数据萃取，反哺舆情数据中台，在线量化形成社情风险指数；第三，推出社情风险指数榜单产品，灵活对时间、地域、属性等维度的社情状况综合评估评判。

3.2 舆情多维交叉比对与可视化

舆情多维交叉比对与可视化是南方舆情“数据组装化、应用服务化”的应用案例。系统操作界面如图 2，它的生成逻辑和技术实现步骤如图 2 所示。



图 2　舆情多维交叉比对

第一，整合全域数据，统一数据出口和查询逻辑，建立舆情态势感知体系，既能对广东地域内开展全面舆情巡查，又能快速发现和展示服务用户以及突发事件的演变趋势；第二，通过复用公共定量指标、加工个性变量指标的方式，既提取领导力、发展力、执行力、创新力等方向指标，又深入到媒体关注、社会维稳、营商环境等细化指标，建立用户坐标系，通过算法模型匹配，迅速清晰地进行用户画像；第三，通过业务应用操作，及时响应和可视化输出数据采集分析和指标模型效果（适应不同呈现载体），通过不同指标数值的阈值设置实现自动预警，快速完成面向用户需求的数据封装和应用服务。

结语

舆情数据中台的核心是数据模型、算法服务和数据产品等能力，通过搭建灵活快速应对变化的架构，更快实现前端产品需求。一方面避免业务高度复用的功能重复建设，另一方面所有业务触点信息均可流向中台，解决数据孤岛，形成信息共享。借助中台的沉淀能力，研发更灵活、业务更敏捷。下一步，舆情数据中台也将阶段性演进，不断形成“技术平台＋建设方法论＋数据产品＋运营服务”解决方案的整体输出，快速调整应对未来的市场变化。媒

参考文献

[1] 钟华，企业 IT 架构转型之道——阿里巴巴中台战略思想和架构实战 [M]. 北京：机械工业出版社，2017.

[2] 喻国明，马思源．人工智能提升网络舆情分析能力 [J]，网络传播，2017（2）：85−87.

[3] 马梅，刘东苏，李慧．基于大数据的网络舆情分析系统模型研究 [J]. 情报科学，2016，36（3）：25−28.

（作者单位：南方报业传媒集团）